THE COCOA GENOME HUB, AN INTEGRATED PLATFORM TO ACCESS THE CRIOLLO GENOME V2

Xavier Argout
Guillaume Martin, Gaetan Droc
Centre de coopération International en Recherche Agronomique pour le Developpement (CIRAD)

Abstract

The first draft genome of the species, from the Belizian Criollo B97-61/B2 cultivar, was published in 2011. Although a useful resource, some improvements were possible, including to identify misassemblies, to reduce the number of scaffolds and gaps, and to anchor un-anchored sequences to the 10 chromosomes. In 2017, we used a combination Next Generation Sequencing data to produce the version 2 of the assembly. We corrected misassembled regions and reduced the number of scaffolds from 4,792 in assembly V1 to 554 in V2 with a N50 increased from 0.47 Mb in V1 to 6.5 Mb in V2. A total of 96.7% of the assembly was anchored to the 10 chromosomes compared to 66.8% in the previous version. Unknown sites (Ns) were reduced from 10.8% to 5.7%. In addition, we updated the functional annotations and performed a new RefSeq structural annotation based on RNAseq evidence. In that context and to support post-genomics efforts, we developed the Cocoa Genome Hub (http://cocoa-genome-hub.southgreen.fr/), an integrated web-based database providing centralized access to T. cacao genome and analysis tools to facilitate basic, translational and applied research in cocoa. We provide access to the complete criollo genome sequence V2 along with gene structure, gene product information, metabolism, gene families, transcriptomics (ESTs, RNA-Seq), genetic markers and genetic maps. The hub relies on generic software (e.g. GMOD tools) for easy querying, visualizing and downloading research data. It includes a Genome Browser enhanced by a Community Annotation System, enabling the improvement of automatic gene annotation through an annotation editor.

Résumé

Le premier brouillon du génome de l'espèce de la variété cultivée Criollo de Belize B97-61 / B2, fut publié en 2011. Bien que ce fût une ressource utile, des améliorations ont été possibles, même pour repérer des erreurs d'assemblage, pour réduire le nombre d'échafaudages et de brèches, et pour ancrer des séquences non ancrées avec les 10 chromosomes. En 2017, nous avons utilisé un mélangé de données de nouvelles technologies de séquençage pour produire la version 2 de l'assemblage. Nous avons corrigé les régions mal assemblées et nous avons réduit le nombre d'échafaudages de 4.792 en V1 à 554 en V2 avec un N50 augmenté de 0.47 Mb en V1 à 6.5 Mb en V2. Un total de 96.7% de l'assemblage s'est ancré sur les 10 chromosomes par rapport au 66.8% dans la version précédente. Les sites inconnus (Ns) ont été réduits de 10.8% à 5.7%. En outre, nous avons mis à jour les notes fonctionnelles et nous avons réalisé une nouvelle note structurelle RefSeq basée sur l'évidence RNAseq. Dans ce contexte, et pour soutenir les efforts postérieurs à la génomique, nous avons développé le Centre du Génome du Cacao (Cocoa Genome Hub) (http://cocoa-genome-hub.southgreen.fr/), une base de données intégrée sur le réseau qui fournit un accès centralisé au génome du T. cacao et des outils d'analyse pour faciliter la recherche basique, translative et appliquée du cacao. Nous fournissons l'accès à la séquence complète du génome Criollo V2 avec la structure des gènes, l'information de produits de gènes, le métabolisme, les familles de gènes, la transcriptomique (EST, ARN-Seq), les marqueurs génétiques et les cartes génétiques. Le centre se base sur un logiciel générique (par exemple, les outils du GMOD) pour consulter, visualiser et télécharger facilement des données de recherche. Il comprend un navigateur du Génome amélioré pour un Système de Notation Communautaire, qui permet l'amélioration de la notation génétique automatique à travers un éditeur de notations.

Resumen

El primer borrador del genoma de la especie, de la variedad cultivada Criollo de Belice B97-61 / B2, se publicó en 2011. Aunque fue un recurso útil, algunas mejoras fueron posibles, incluso para identificar errores en el ensamblaje, para reducir el número de andamios y brechas, y para anclar secuencias no ancladas a los 10 cromosomas. En 2017, utilizamos una combinación de datos de nuevas tecnologías secuenciación para producir la versión 2 del ensamblaje. Corregimos las regiones mal ensambladas y redujimos la cantidad de andamios de 4.792 en V1 a 554 en V2 con un N50 aumentado de 0.47 Mb en V1 a 6.5 Mb en V2. Un total del 96.7% del ensamblaje se ancló a los 10 cromosomas en comparación con el 66.8% en la versión anterior. Los sitios desconocidos (Ns) se redujeron de 10.8% a 5.7%. Además, actualizamos las anotaciones funcionales y realizamos una nueva anotación estructural RefSeq basada en la evidencia RNAseq. En ese contexto y para apoyar los esfuerzos posteriores a la genómica, desarrollamos el Centro de Genoma del Cacao (Cocoa Genome Hub) (http://cocoa-genome-hub.southgreen.fr/), una base de datos integrada basada en la red que proporciona acceso centralizado al genoma del T. cacao y herramientas de análisis para facilitar la investigación básica, traslativa y aplicada en el cacao. Proporcionamos acceso a la secuencia completa del genoma criollo V2 junto con la estructura de genes, información de productos de genes, metabolismo, familias de genes, transcriptómica (EST, ARN-Seq), marcadores genéticos y mapas genéticos. El centro se basa en un software genérico (por ejemplo, las herramientas de GMOD) para consultar, visualizar y descargar fácilmente datos de investigación. Incluye un navegador de Genoma mejorado por un Sistema de Anotación Comunitaria, que permite la mejora de la anotación genética automática a través de un editor de anotaciones.

International Symposium on Cocoa Research 2017

BOOKLET OF ABSTRACTS

LIVRET DES RÉSUMÉS

FOLLETO DE RESUMENES

13-17 November 2017, Swissôtel, Lima, Peru

INTERNATIONAL COCOA ORGANIZATION

icco.org/iscr2017

International Symposium on Cocoa Research

icco.org/iscr2017